Predicting the protein folding nucleus from a sequence

Anne Poupon[a,b,*], Jean-Paul Mornon[a]

[a] *Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS UMR C7590, Universités P6 et P7, T16, Case 115, 4, place Jussieu, 75232 Paris Cedex 05, France*
[b] *Département d'Ingénierie et d'Etude des Protéines, Bat. 152, CEA/SACLAY, 91 191 Gif-sur-Yvette Cedex, France*

Received 25 March 1999

Abstract Understanding the mechanism of protein folding would allow prediction of the three-dimensional structure from sequence data alone. It has been shown that small proteins fold in a small number of kinetic steps and that significantly populated intermediate states exist for some of them. Studies of these intermediates have demonstrated the existence of specific interactions established during the initial stages of folding. Comparison of the amino acids participating in these specific and essential interactions and constituting the folding nucleus with conserved hydrophobic positions of a given fold shows a striking correspondence. This finding opens the perspective of predicting the folding nucleus knowing only a set of divergent sequences of a protein family.

Key words: Protein folding; Folding intermediate; Hydrophobic packing; Sequence similarity

1. Introduction

Consequences of mutations in proteins are very difficult to predict. Indeed, some mutations, even numerous, have no effect on the biological activity and some point mutations, affecting amino acids far from the active site, completely abolish activity. A similar problem is encountered with mutant proteins that cannot be folded, particularly through chemical synthesis. These features cannot generally be explained by changes in the protein stability, but rather by drastic perturbations of the folding pathway. Understanding of this phenomenon would therefore require a better knowledge of folding mechanisms.

Study of the early stages of protein folding is a considerable task. Their short lifetimes have necessitated the development of very specific and fastidious techniques requiring the use of appreciable amounts of pure protein [1–11] and, sometimes, the construction of many mutants [2,12,13]. The structure and particularities of these transition states can also be investigated by different computational techniques, simulation techniques such as molecular dynamics [13–15] or lattice model simulations [16,17]. Much time could be saved if it was possible to predict, from sequences only, a panel of amino acids that are essential in the folding process.

The study of topohydrophobic positions in fold families [18,19] (positions occupied by hydrophobic amino acids in each member of the family) has shown that the residues occupying these positions have very distinct properties. They are very significantly more buried than the hydrophobic amino acids in non-topohydrophobic positions and are also less dispersed (the axis between Cα and the gravity center of an amino acid's side chain in a topohydrophobic position is more conserved upon mutation than that of a hydrophobic amino acid in a non-topohydrophobic position). They form a lattice of interacting residues in the inner core of the protein. These findings make them good candidates for involvement in the folding nucleus.

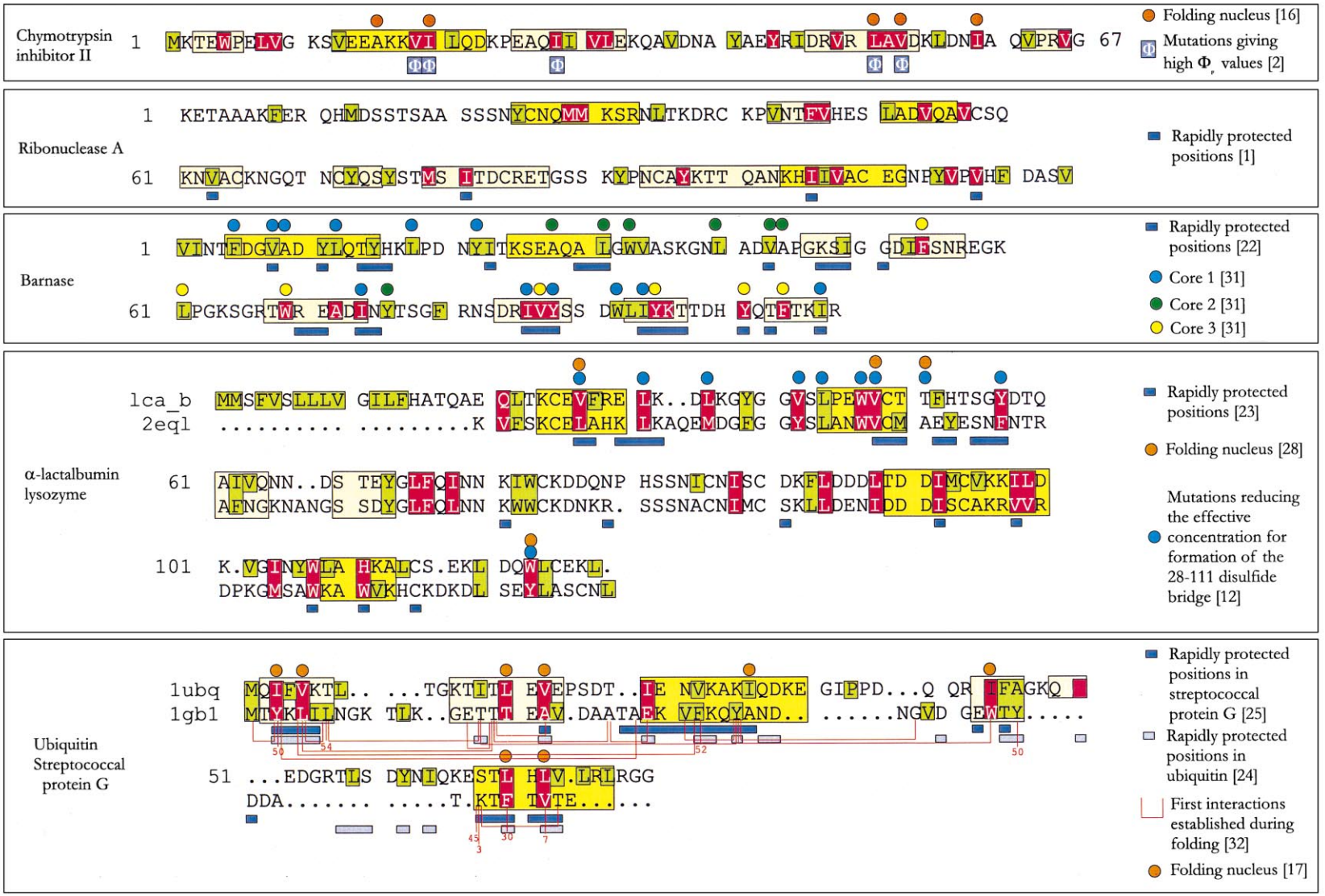
For each protein in which folding was studied (indicated below), topohydrophobic positions were determined and compared with the amino acids identified experimentally or by simulation (Table 1 and Fig. 1). Topohydrophobic positions were determined from multiple alignments including all the proteins with a known three-dimensional (3D) structure sharing the same fold, together with divergent sequences showing a significant sequence identity with members of the family, with elimination of the strongest redundancies (for any pair of proteins in a family, the sequence identity is lower than 55%). Protein sequences were aligned on the basis of the structure when available [19] or by sequence comparison using the sensitive two-dimensional (2D) HCA method [20,21]. Topohydrophobic positions are positions within the multiple alignment containing at least 75% strong hydrophobic amino acids (VILFMYW) and where the remaining amino acids have good propensities for regular secondary structures and belong to the ACTQERKH group [21] (mainly ACT).

2. Results and discussion

Different types of studies have been performed to explore the mechanism of folding. Protection of the amide protons during folding can be measured by ^{1}H-NMR, as shown for ribonuclease A [1], barnase [22], lysozyme [23], ubiquitin [24], streptococcal protein G [25], apomyoglobin [8], cytochrome *c* [4] and ribonuclease H [26]. Protein engineering studies can determine how each residue contributes to the stabilization of the intermediate state by mutating it to alanine and/or glycine. Fractional Φ values (Φ_F), which estimate the extent of interactions of each residue at different stages in the folding pathway [27] (this has been performed for chymotrypsin inhibitor II [2]), can be computed. Other folding parameters can also be measured, such as the effective concentration for formation of the 28-111 disulfide bond in α-lactalbumin [12,28]. In the case of calmodulin, it has even been shown that the mutant proteins do not fold completely in the absence of calcium [29]. Formation of H-bonds during folding can be assessed by H exchange pulse-labelling (apomyoglobin [30]) or by molecular dynamics simulations (chymotrypsin inhibitor II [14], barnase [31], streptococcal protein G [32]). Finally, lattice model simulation methods generate random sequences whose folding on a cubic lattice is simulated. The fast folding sequences are compared with each other (chymotrypsin inhibitor II [16], ubiquitin [17]).

*Corresponding author. Fax: (33) (1) 69 08 90 71.
E-mail: apoupon@cea.fr

0014-5793/99/$20.00
PII: S0014-5793(99)00622-5


Chymotrypsin inhibitor II
Folding nucleus [16]
Mutations giving high Φ, values [2]
Ribonuclease A
Rapidly protected positions [1]
Barnase
Rapidly protected positions [22]
Core 1 [31]
Core 2 [31]
Core 3 [31]
α-lactalbumin lysozyme
Rapidly protected positions [23]
Folding nucleus [28]
Mutations reducing the effective concentration for formation of the 28-111 disulfide bridge [12]
Ubiquitin Streptococcal protein G
Rapidly protected positions in streptococcal protein G [25]
Rapidly protected positions in ubiquitin [24]
First interactions established during folding [32]
Folding nucleus [17]

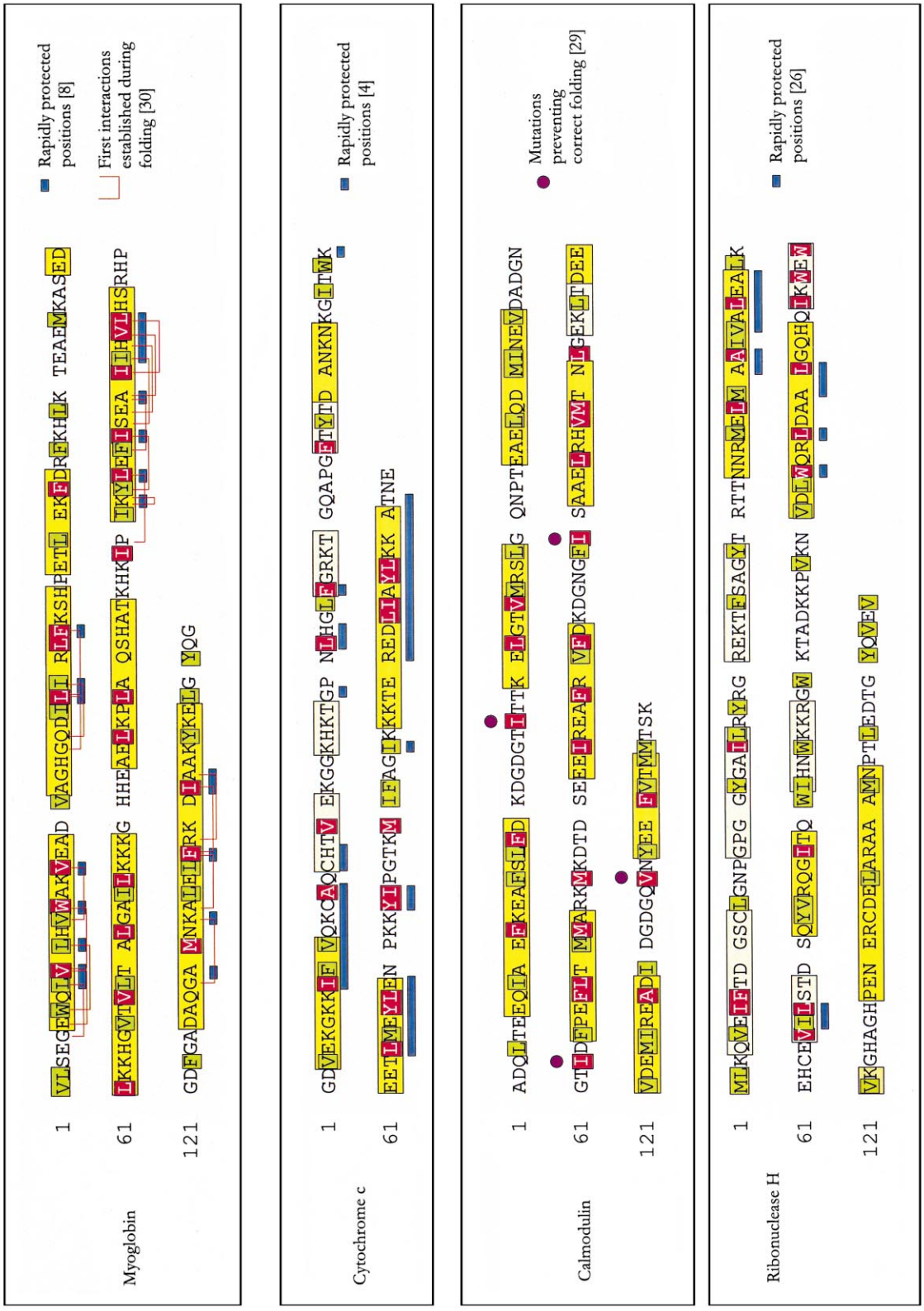


Fig. 1. For each protein considered in this paper (Table 1), the sequence is shown with regular secondary structures on a yellow background (dark for helices, light for strands, assignments made with P_SEA [35]). Topohydrophobic positions are white on a red background, hydrophobic amino acids in non-topohydrophobic positions are shown on a green background. Results of the different studies are shown above or below the sequences.

Table 1
Experimental and simulation data, comparison with topohydrophobic positions

Protein	Methods	Results	Other members	NT	Comparison	*R*
Chymotrypsin inhibitor II	Lattice model simulations [16]	Complete determination of the folding nucleus	PDB: 1tin 2sec_i 1cis	9 (24)	Four hydrophobic residues constitute the folding nucleus, all in topohydrophobic positions	+
Id: 40.8% RMS: 2.5			SW: ici1_lycpe, ier1_lyces			
	Molecular dynamics simulations [14]	Complete determination of the folding nucleus				
	Protein engineering [2]	Computation of ΦF for mutants, estimation of the number of native-like contacts in the transition state for each mutated residue			Five hydrophobic positions have high ΦF values, all in topohydrophobic positions	+
Ribonuclease A	2D ^{1}H-NMR [1]	Protection factors of eight residues (three of them unidentified) in an early folding intermediate	PDB: 1ang 1onc 1rbc	14 (27)	Four hydrophobic residues are rapidly protected, three are in topohydrophobic positions	3.3
Id: 41.6% RMS: 1.4			SW: ecp_rat lecs_ranja			
Barnase	Molecular dynamics simulations [31]	Analysis of the unfolding of barnase, observation of the solvation of the hydrophobic cores	PDB: 1bse_a 1gmp_a 1rnt 1aqz_a	11 (35)	Core 3 is the last to unfold, six of the seven residues of this core are in topohydrophobic positions	13.6
Id: 21.3% RMS: 2.8						
	2D ^{1}H-NMR [22]	Protection factors of some residues in partly unfolded states			16 Hydrophobic residues are rapidly protected, seven are in topohydrophobic positions	1.7
α-Lactalbumin	Protein engineering [12]	Effects on the folding rate of alanine substitution of buried amino acids measured by the oxidation rate of the 28-111 disulfide bridge	PDB: 2ihl 2eql 1hml	21 (42)	10 Hydrophobic/alanine substitutions reduce the effective concentration for formation of the 28-111 disulfide bridge by more than 40%, eight of them are in topohydrophobic positions	4
Id: 31.8% RMS: 1.5			SW: lca_bovin lyc1_tacac ipgf_shifl iagb_salti x193_ecoli			
	Protein engineering [28]	Effects on the folding rate of alanine substitution of buried amino acids measured by the oxidation rate of the 28-111 disulfide bridge			Four hydrophobic amino acids identified in the folding nucleus, three are in topohydrophobic positions	3
Lysozyme	H exchange pulse-labelling [23]	Time courses of the change in proton occupancies during folding		(34)	15 Hydrophobic amino acids are rapidly protected, 12 are in topohydrophobic positions	3.5
Ubiquitin	2D ^{1}H-NMR [24]	Protection factors of each residue	PDB: 1frd 1frr_a 2pia 1put 1ubq 1gua 1alo 1ste 1esf_b 2qil_a 1gb1	10 (24)	16 Hydrophobic residues are protected in the intermediate state, eight are in topohydrophobic positions	1.7
Id: 26.4% RMS: 2.7						
	Lattice model simulations [17]	Identification of seven putative folding nucleus residues			Identification of seven potential residues of the folding nucleus, six are in topohydrophobic positions	10.2

Table 1
Experimental and simulation data, comparison with topohydrophobic positions

Protein	Methods	Results	Other members	NT	Comparison	*R*
Streptococcal protein G	Molecular dynamics simulations [32]	Chronology of the establishment of the intramolecular interactions during folding		(19)	13 Hydrophobic positions rapidly establish interactions, six are in topohydrophobic positions	1, 2.3[a]
	Quenched-flow D-H experiments [25]	Protection factors of residues implicated in the interaction between the secondary structures			10 Hydrophobic residues are rapidly protected, seven are in topohydrophobic positions	2.1
Apomyoglobin	2D ^{1}H-NMR [8]	Protection factors of 39 residues in partly unfolded states	PDB: 1myt 1myc 1dxt_a 1dxt_b1pbx_a 1pbx_b 1hlb 1ith_a 1eca 1flp 1mba 2lhb 1hbg 1gdi 1ash	22 (48)	15 Hydrophobic residues are rapidly protected, 11 are in topohydrophobic positions	3.2
Id: 22.8% RMS: 2.4						
	H exchange pulse-labelling [30]	Kinetics of formation of the hydrogen bonds during folding			15 Hydrophobic residues establish their interactions within 5 ms of folding, 11 are in topohydrophobic positions	3.2, +[a]
Cytochrome *c*	2D ^{1}H-NMR [4]	Protection factors of each residue	PDB: 1cri 1cry 1cot 3c2c	16 (27)	17 Hydrophobic residues are rapidly protected, 13 are in topohydrophobic positions	3.6
Id: 34.1% RMS: 1.4			SW: cyc_horse cox2_bacp3 cox2_bacfi			
Calmodulin	Protein engineering [29]	Mutation of four residues essential for folding	PDB: 2scp_a 2bbm_a 2sas 1rro 2pal 1rtp_1 1rec	21 (44)	The four mutated residues are in topohydrophobic positions	+
Id: 26.4% RMS: 2.0						
Ribonuclease H	H exchange pulse-labelling [26]	Protection factors of some residues in an early folding intermediate	PDB: 2rn2 1hrh	15 (45)	Eight hydrophobic residues are rapidly protected, six are in topohydrophobic positions	6.1
Id: 24.2%			SW: rnh_helpy ypdq_bacsu rnh1_yeast rnh1_crifa pol_sfv3l pol3_baevm pol_srv2 pol_hv2g1			
RMS: 1.9						
Mean values	Id: $m = 30\%$, $\delta = 7.8\%$	RMS: $m = 2.1$, $\delta = 0.6$		x, $m = 0.45$, $\delta = 0.09$	y, $m = 0.78$, $\delta = 0.21$	

Protein: protein studied (Id: mean sequence identity; RMS: mean root mean square deviation between the 3D structures of the family); Methods: methods used for the study and bibliographic reference; Results: information deduced for the corresponding protein; Other members: other structures (identified by their PDB codes [33]) and sequences (identified by their Swiss-Prot code [34]) used for multiple alignment; NT: number of topohydrophobic positions in the family, total number of hydrophobic amino acids between brackets; Comparison: comparison between experimental or simulation results and topohydrophobic positions; *R*: coincidence ratio (as defined in the text). Last line: mean values (m) and standard deviation (δ): Id, mean sequence identity; RMS: mean root mean square deviation; x: mean proportion of hydrophobic residues in topohydrophobic positions in the sequence; y: mean proportion of hydrophobic residues in topohydrophobic positions in the identified subset (folding nucleus, rapidly protected residues).

[a]In the study of early interactions in apomyoglobin [30], the identified interactions are clearly not independent of each other. Four 'groups' can be defined (group I: 5-9, 6-10, 7-11; group II: 25-29, 26-30; group III: 106-110, 108-112, 109-113, 110-114, 111-115; group IV: 138-142, 139-143, see Fig. 1). In each of these groups, at least one interaction involves a residue in a topohydrophobic position. Considering one interaction as representative of the whole group, the coincidence ratio *R* is infinite. Similarly, in the case of streptococcal protein G [32] (which is the only case of a coincidence ratio equal to one), there is one group (5-16, 6-15, 7-14) and the first interaction established (5-16) involves one residue in a topohydrophobic position. By considering this group as a single interaction, the coincidence ratio *R* increases to 2.3.

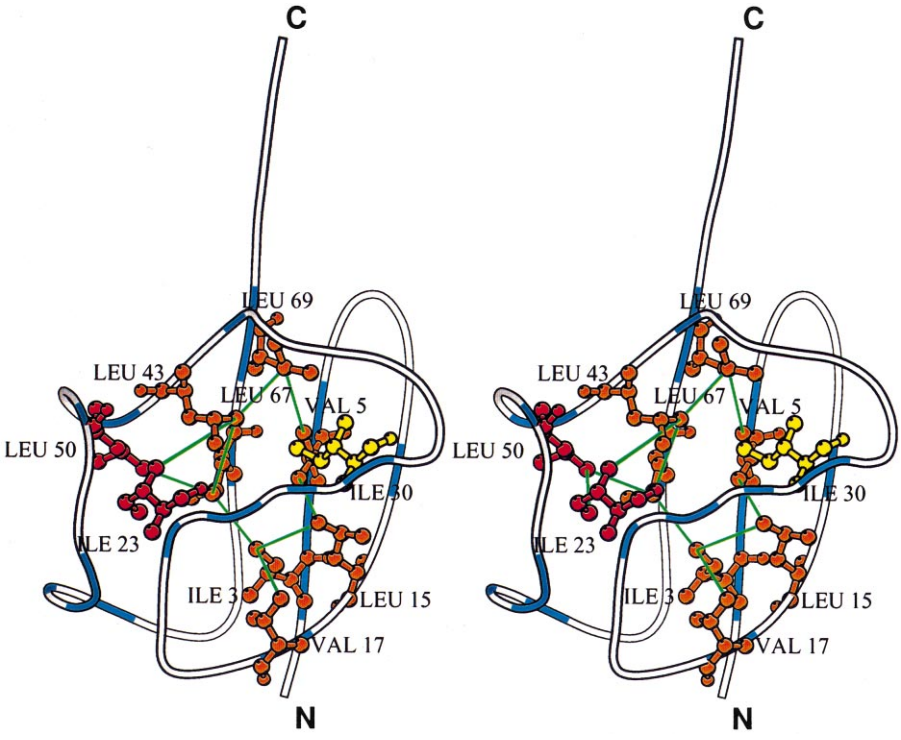


Fig. 2. Stereographic view of ubiquitin side chains of amino acids in topohydrophobic positions are depicted in a ball and stick representation and colored orange for amino acids which also constitute the folding nucleus. I-30 which belongs to the folding nucleus but is not a topohydrophobic position is indicated in yellow. Inversely, I-23 and L-50 which are topohydrophobic positions but not folding nucleus units are colored red. The backbone of the protein is colored in blue in rapidly protected regions [24]. The network of interacting residues formed by the amino acids in topohydrophobic positions [18,19] is figured with green lines.

For each protein studied, a coincidence ratio R was computed (Table 1). This ratio estimates the preference of the studied property for hydrophobic amino acids in topohydrophobic positions versus hydrophobic amino acids in non-topohydrophobic positions. This ratio is defined according to the hypothesis that, for a given property, the probability to be positive for a topohydrophobic residue T compared to the same probability for a non-topohydrophobic residue NT is: $P(\mathrm{T}) = \rho \cdot P(\mathrm{NT})$. If the property studied is specific to hydrophobic amino acids in non-topohydrophobic positions, ρ is equal to 0. If the property is indifferent to the topohydrophobic nature of the amino acids, ρ is equal to 1. If the property is specific to hydrophobic amino acids in topohydrophobic positions, ρ tends to infinity. This definition implies that ρ can be estimated by computing, for each experiment, the ratio R (ρ being the theoretical value of the preference, it cannot be calculated directly from the experiment):

$$R = \frac{y(1-x)}{x(1-y)}$$

where x and y are the proportions of hydrophobic residues in topohydrophobic positions in the sequence and in the subset identified by the experiment, respectively. For example, in the sequence of apomyoglobin, there are 48 hydrophobic amino acids, 22 of which are in topohydrophobic positions. Therefore, x is equal to 0.46. In this protein, 15 hydrophobic amino acids are rapidly protected during folding, 11 of which are in topohydrophobic positions. Therefore, y is equal to 0.73. Using these values, $R = 3.2$.

This ratio is always greater than one for all examples but one, meaning that for any property studied in the intermediate states or in the initial stages of folding, there is a clear preference for hydrophobic amino acids in topohydrophobic positions.

For proton protection studies, the ratio is generally high, with a mean logarithmic value of 2.8 (a logarithmic mean value was used because R is a multiplicative factor). However, two of these values are quite low, 1.74 for barnase and 1.29 for ubiquitin. In the case of barnase, this might be due to the low number of sequences in the multiple alignment (only four) leading to an overestimation of topohydrophobic positions [18,19]. Conversely, in the case of ubiquitin, this poor result might be due to the high structural divergence in the family. Indeed, we have previously shown that the correct identification of topohydrophobic positions requires the multiple alignment of six or more divergent sequences, but the corresponding proteins must be structurally homogeneous [18,19].

Protein engineering studies give a very high coincidence ratio, two infinite values (in the two studies considered, all the hydrophobic residues for which mutation influences the early stages of folding are in topohydrophobic positions) and two other values equal to three and four, which indicates a clear preference for hydrophobic amino acids in topohydrophobic positions. Finally, the studies giving the highest coincidence ratio are the simulations (two infinite and 7.87). In two individual cases, chymotrypsin inhibitor II [16] and ubiquitin [17], the study of a large number of fast folding artifical sequences [16], or the comparison of different members of a family [17], has already revealed the relationships between the folding nucleus and conserved hydrophobic amino acids.

In conclusion, what was previously suggested in two indi-

vidual cases [16,17] appears to be a general feature of globular proteins. Most of the hydrophobic amino acids belonging to the folding nucleus, or rapidly protected during folding or playing an important role in the first stages of folding, occupy positions that we have defined as topohydrophobic (Fig. 2). As a result, most of these amino acids can be predicted from the sequence alone and determined from accurate alignments of a limited set of divergent sequences. Moreover, the close contact lattice of topohydrophobic positions would bring useful geometric constraints to the emerging procedures of ab initio protein folding predictions and therefore, may be of vital importance.

This prediction could also allow a better prediction of mutation effects. Indeed, mutation of a residue in a topohydrophobic position by a non-hydrophobic residue could alter the folding of the protein. Some examples of such mutations were previously commented for calmodulin [29].

References

[1] Udgaonkar, J.B. (1988) Nature (London) 335, 694–699.
[2] Itzhaki, L.S., Otzen, D. and Fersht, A.R. (1995) J. Mol. Biol. 254, 260–288.
[3] Huang, G.S. and Oas, T.G. (1995) Biochemistry 34, 3884–3892.
[4] Sosnick, T.R., Mayne, L. and Englander, S.W. (1996) Proteins 24, 413–426.
[5] Jackson, S.E. (1998) Fold. Des. 3, R81–R91.
[6] Roder, H., Elšve, G.A. and Englander, S.W. (1988) Nature (London) 335, 700–704.
[7] Bashford, D., Cohen, F.E., Karplus, M. and Weaver, D.L. (1988) Proteins 4, 211–227.
[8] Hughson, F.M., Wright, P.E. and Baldwin, R.L. (1990) Science 249, 1544–1548.
[9] Fersht, A.R. (1997) Curr. Opin. Struct. Biol. 7, 3–9.
[10] Baldwin, R.L. and Rose, G.D. (1999) TIBS 24, 77–83.
[11] Killick, T.R., Freund, S.M.V. and Fersht, A.R. (1998) FEBS Lett. 423, 110–112.
[12] Song, J., Bai, P., Luo, L. and Peng, Z.J. (1998) Mol. Biol. 280, 167–174.
[13] Dagget, V., Li, A., Itzhaki, L.S., Otzen, D.E. and Fersht, A.R. (1996) J. Mol. Biol. 257, 430–440.
[14] Li, A. and Daggett, V. (1996) J. Mol. Biol. 257, 412–429.
[15] Alonso, D.O.V. and Daggett, V.J. (1995) Mol. Biol. 247, 501–520.
[16] Shakhnovich, E., Abkevich, V. and Ptitsyn, O. (1996) Nature (London) 379, 96–98.
[17] Michnik, S.W. and Shakhnovich, E. (1998) Fold. Des. 3, 239–251.
[18] Poupon, A. and Mornon, J.-P. (1999) Theor. Chem. Acc. 101, 2–8.
[19] Poupon, A. and Mornon, J.-P. (1998) Proteins 33, 329–342.
[20] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.-P. (1987) FEBS Lett. 224, 149–155.
[21] Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. and Mornon, J.-P. (1997) Cell. Mol. Life Sci. 53, 621–645.
[22] Bycroft, M., Matouschek, A., Kellis, J.T., Serrano, L. and Fersht, A.R. (1990) Nature (London) 346, 488–490.
[23] Radford, S.E., Dobson, C.M. and Evans, P.A. (1992) Nature (London) 358, 302–307.
[24] Pan, Y. and Briggs, M.S. (1992) Biochemistry 31, 11405–11412.
[25] Kuszewski, J., Clore, G.M. and Gronenborg, A.M. (1994) Protein Sci. 3, 1945–1952.
[26] Raschke, T.M. and Marqusee, S. (1997) Nat. Struct. Biol. 4, 298–304.
[27] Fersht, A.R. (1995) Curr. Opin. Struct. Biol. 5, 79–84.
[28] Wu, L.C. and Kim, P.S. (1998) J. Mol. Biol. 280, 175–182.
[29] Browne, J.P., Strom, M., Martin, S.R. and Bayley, P.M. (1997) Biochemistry 36, 9550–9561.
[30] Jennings, P.A. and Wright, P.R. (1993) Science 262, 892–896.
[31] Caflisch, A. and Karplus, M.J. (1995) Mol. Biol. 252, 672–708.
[32] Sheinerman, F.B. and Brooks, C.L. (1998) J. Mol. Biol. 278, 438–456.
[33] Bernstein, F.C., Koetzle, T.F. and Williams, G.J. (1977) Mol. Biol. 112, 535–542.
[34] Bairoch, A. and Apweiler, R. (1998) Nucleic Acid Res. 26, 38–42.
[35] Labesse, G., Colloc'h, N., Pothier, J. and Mornon, J.-P. (1997) CABIOS 13, 291–295.